Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713428038

On implementation of the Gibbs sampler for estimating the accuracy of multiple diagnostic tests

Fabio Principato[a]; Angela Vullo[b]; Domenica Matranga[c]
[a] Dipartimento di Fisica e Tecnologie Relative, Universitá di Palermo, Palermo, Italy [b] Area Sorveglianza Epidemiologica, Istituto Zooprofilattico Sperimentale della Sicilia 'A.Mirri', Palermo, Italy [c] Dipartimento di Biotecnologie Mediche e Medicina Legale, Sezione Scienze Radiologiche Universitá di Palermo, Palermo, Italy

Online publication date: 11 August 2010

To cite this Article Principato, Fabio , Vullo, Angela and Matranga, Domenica(2010) 'On implementation of the Gibbs sampler for estimating the accuracy of multiple diagnostic tests', Journal of Applied Statistics, 37: 8, 1335 — 1354
To link to this Article: DOI: 10.1080/02664760903030239
URL: http://dx.doi.org/10.1080/02664760903030239

Journal of Applied Statistics
Vol. 37, No. 8, August 2010, 1335–1354

Taylor & Francis
Taylor & Francis Group

On implementation of the Gibbs sampler for estimating the accuracy of multiple diagnostic tests

Fabio Principato[a]*, Angela Vullo[b] and Domenica Matranga[c]

[a]Dipartimento di Fisica e Tecnologie Relative, Universitá di Palermo, Viale delle Scienze, Edificio 18, 90128 Palermo, Italy; [b]Area Sorveglianza Epidemiologica, Istituto Zooprofilattico Sperimentale della Sicilia 'A.Mirri', Palermo, Italy; [c]Dipartimento di Biotecnologie Mediche e Medicina Legale, Sezione Scienze Radiologiche Universitá di Palermo, Palermo, Italy

(*Received 6 August 2008; final version received 6 May 2009*)

Implementation of the Gibbs sampler for estimating the accuracy of multiple binary diagnostic tests in one population has been investigated. This method, proposed by Joseph, Gyorkos and Coupal, makes use of a Bayesian approach and is used in the absence of a gold standard to estimate the prevalence, the sensitivity and specificity of medical diagnostic tests. The expressions that allow this method to be implemented for an arbitrary number of tests are given. By using the convergence diagnostics procedure of Raftery and Lewis, the relation between the number of iterations of Gibbs sampling and the precision of the estimated quantiles of the posterior distributions is derived. An example concerning a data set of gastro-esophageal reflux disease patients collected to evaluate the accuracy of the water siphon test compared with 24 h pH-monitoring, endoscopy and histology tests is presented. The main message that emerges from our analysis is that implementation of the Gibbs sampler to estimate the parameters of multiple binary diagnostic tests can be critical and convergence diagnostic is advised for this method. The factors which affect the convergence of the chains to the posterior distributions and those that influence the precision of their quantiles are analyzed.

Keywords: Gibbs sampler; Bayesian analysis; convergence diagnostics; diagnostic tests; gastro-esophageal reflux disease

1. Introduction

Diagnostic tests based on clinical observations or measurements of biological quantities relevant to the infection status of a subject are frequently used in epidemiology. Diagnostic tests are also important in medicine because they form the basis of screening programs for early diagnoses of diseases. However, tests are imperfect tools because healthy subjects are often mistakenly classified as diseased while truly diseased subjects are classified as non-diseased.

*Corresponding author. Email: principato@unipa.it

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760903030239
http://www.informaworld.com

A common practice in many diagnostic accuracy studies is to evaluate a novel test by using an imperfect standard as if it were a gold standard. The effect of this is to obtain strongly biased test accuracy estimates: if the test and the imperfect gold standard are conditionally independent, then the sensitivity and the specificity of the new test will be underestimated. By contrast, if they are highly correlated, given the condition status, then the test and the gold standard will tend to misclassify the same patients and the accuracy of the tests will be overestimated accordingly [25].

In order to correct imperfect standard bias, the latent class analysis approach has been proposed as a probabilistic model for the relation between the novel diagnostic test, one or more imperfect reference tests and the latent, unobserved disease status [4,26]. However, there are some problems with this method: firstly, the method can proceed without any clinical definition of the disease, but prevalence and test accuracy parameters have a doubtful clinical meaning; secondly, severe bias occurs when the conditional independence assumption is violated; finally, the estimation algorithm can converge to a local maximum solution and strategies must be adopted to avoid this [20].

As an alternative, the Bayesian method has been proposed to estimate accuracy parameters of the tests by first imposing a prior distribution over all unknown parameters [15]. Bayesian inference about a parameter is made using the posterior distribution which is computed by combining the likelihood function of the observed data with the prior distribution. The advantage of Bayesian analysis for diagnostic test evaluation is to incorporate prior scientific information about the sensitivities and specificities of the tests and prior information about the prevalence of the sampled populations.

The Bayesian approach considers uncertainties associated with all unknown quantities, whether they are observed or unobserved. Inference is drawn by constructing the joint probability distribution of all unobserved quantities based on everything that is known about them. This knowledge incorporates previous information about the phenomena under study and is also based on values of observed quantities, when available. In this case, the distribution of the unknown but observable data is called the posterior distribution because it is obtained after the data have been observed [8]. In [15] the Gibbs sampler was used to find the parameters of two diagnostic tests and one sample of data without a gold standard. In [14] this method was used to estimate the prevalence of osteoarthritis in three diagnostic tests. The method has the main advantage of drawing inference from diagnostic tests in the absence of a gold standard. User-friendly software implementing the Gibbs sampler of [15] for up to three diagnostic test data is available from the web page [14]. However, the computational aspects of implementing the algorithm in this method are scarcely considered.

As with all Markov chain Monte Carlo (MCMC) methods, the Gibbs sampler implementation may be complicated by the potential difficulty of quantitatively assessing convergence of the Markov chains to target distributions, for which purpose several methods have been proposed for convergence diagnostics [3]. Raftery and Lewis (R&L) [21,22] provide a method which allows the number of iterations of the Gibbs sampler required to estimate the quantiles of the posterior distribution to be determined in advance. This method, which applies more generally to MCMC schemes besides the Gibbs sampler, also determines the spacing between iterations retained for the final analysis and the number of initial burn-in iterations discarded. The number of iterations after the burn-in stage depends on the precision of the estimates of quantiles of interest of the posterior distribution, which has to be known.

The objective of this work is to assess computational implementation of the Gibbs sampler in the method proposed in [15], used to estimate the parameters of diagnostic tests without a gold standard, by using the technique proposed by R&L. For this purpose, we first generalize the procedure proposed in [15] to the case of N diagnostic tests and one population by writing the analytical expressions of the model. We then consider the question relating to the choice of the number of iterations of the Gibbs sampler needed to summarize the information through the posterior medians and their credibility intervals with known precision. The relation between the number of iterations and the precision of the estimated quantiles, which depends on the form of the posterior distribution, is derived for the case of the beta distribution, which is the target distribution of the Gibbs sampler for the considered model.

As an example of the procedure proposed here, we evaluate the accuracy of four diagnostic tests used in the case of the gastro-esophageal reflux disease (GERD). In particular, we evaluate the accuracy of the water-siphon test (WST) associated with a barium study compared with another three frequently used imperfect reference tests, i.e. 24 h pH monitoring, endoscopy and histology.

We present the extension of the method proposed in [15] to N diagnostic tests and one population in Section 2. We determine the number of burn-in and sub-sequential iterations needed to obtain the estimate of the medians and 95% of credibility intervals of the diagnostic test parameters with a known precision in Section 3. In Section 4, we introduce an example of a data set of GERD patients collected to evaluate the accuracy of the WST compared with other, more extensively investigated 24 h pH monitoring, endoscopy and histology tests. Finally, the results are discussed in Section 5.

2. Gibbs sampler to estimate the parameters of multiple binary tests in a single population

In this section, we consider the model proposed by Joseph *et al.* for Bayesian estimation of the parameters of diagnostic tests, presented in [15] for two diagnostic tests and applied in [16] for three diagnostic tests. Here, we give the expressions valid for an arbitrary number N of diagnostic binary tests in one population.

Let $t_n = 1$ and $t_n = 0$ denote the positive and negative result, respectively, from the test n, with $n = 1, \ldots, N$ and let D be the true status of the subject, with $D = 1$ and $D = 0$ when the subject is diseased and non-diseased, respectively. With N tests there are 2^N different test results. Denoting with $I_N = \{1, \ldots, 2^N\}$ the set of the index test patterns, the ith result of the tests is indicated with the vector $\mathbf{t}^i = (t_1^i, \ldots, t_N^i)$, with $i \in I_N$. With this notation $\mathbf{t}^i$ is the binary digit of the number $i - 1$. For example, for $N = 4$ we have 16 different test patterns and for $i = 3$ we find that $\mathbf{t}^3 = (0, 0, 1, 0)$.

The sensitivity of the nth test is defined as $Se_n = P(t_n = 1|D = 1)$ and its specificity is $Sp_n = P(t_n = 0|D = 0)$, where $P(A|B)$ denotes the conditional probability of A given B. Let π denote the disease prevalence of the population. Let **Se** and **Sp** denote the vectors with N components of the sensitivity and the specificity of the N tests, respectively.

Let O_i denote the number of subjects that shows the test pattern $\mathbf{t}^i$. If N_s is the number of subjects in the population under observation, then

$$N_\mathrm{s} = \sum_{i \in I_N} O_i.$$

With Y_i, with $i \in I_N$, we denote the number of truly diseased subjects that shows the test pattern as $\mathbf{t}^i$. Hence, Y_i is the latent variable when there is no gold standard. With **Y** and **O** we indicate the vectors with 2^N components of the latent and observed data, respectively.

Let p_i denote the probability that a subject is diseased assuming that shows the ith test pattern, i.e. $P(D = 1|\mathbf{t}^i)$. From Bayes theorem, we obtain

$$p_i = \frac{P(\mathbf{t}^i|D = 1)P(D = 1)}{P(\mathbf{t}^i)}. \tag{1}$$

Here, we assume that the test outcomes for a given subject are independently conditional on the disease status of the subject. Hence, the probability that one subject shows the test pattern $\mathbf{t}^i$

assuming it is diseased is

$$P(\mathbf{t}^i|D=1)=\prod_{n=1}^{N}(Se_n^{t_n^i}(1-Se_n)^{(1-t_n^i)}), \tag{2}$$

and the probability that one subject shows the test pattern $\mathbf{t}^i$ regardless of its true status is

$$P(\mathbf{t}^i)=\pi\prod_{n=1}^{N}(Se_n^{t_n^i}(1-Se_n)^{(1-t_n^i)})+(1-\pi)\prod_{n=1}^{N}(Sp_n^{(1-t_n^i)}(1-Sp_n)^{t_n^i}). \tag{3}$$

Noting that $P(D=1)=\pi$ and substituting Equations (2) and (3) in Equation (1) we obtain

$$p_i=\frac{\pi\prod_{n=1}^{N}(Se_n^{t_n^i}(1-Se_n)^{(1-t_n^i)})}{\pi\prod_{n=1}^{N}(Se_n^{t_n^i}(1-Se_n)^{(1-t_n^i)})+(1-\pi)\prod_{n=1}^{N}(Sp_n^{(1-t_n^i)}(1-Sp_n)^{t_n^i})}. \tag{4}$$

Each latent variable Y_i representing the number of true positive subjects out of O_i in the ith category of the observed test results follows the Binomial distribution

$$Y_i\sim\text{Binomial}(O_i,p_i). \tag{5}$$

The likelihood function over the observed and latent data from the N test model is indicated with $l(\mathbf{O},\mathbf{Y}|\pi,\mathbf{Se},\mathbf{Sp})$ and is constructed by generalizing the procedure present in [15]. More specifically, for each test result $\mathbf{t}^i$ the subject can be either diseased or non-diseased, so there are 2^{N+1} possible combinations of observed and latent data. For the ith test result, the contribution to the likelihood is $\pi\prod_{n=1}^{N}Se_n^{t_n^i}(1-Se_n)^{(1-t_n^i)}$ if the individual is truly infected, otherwise it is $(1-\pi)\prod_{n=1}^{N}(1-Sp_n)^{t_n^i}Sp_n^{(1-t_n^i)}$. In order to write the expression of the likelihood function, we introduce the following set of indices relating to the nth test

$$C_n=\{i\in I_N:t_n^i=1\}\text{ and }\bar{C_n}=\{i\in I_N:t_n^i=0\},$$

with $I_N=C_n\cup\bar{C_n}$. With this set we can write the likelihood function of the augmented data $(\mathbf{O},\mathbf{Y})$ for the N test model

$$l(\mathbf{O},\mathbf{Y}|\pi,\mathbf{Se},\mathbf{Sp})=\pi^{\sum_{i\in I_N}Y_i}(1-\pi)^{(N_s-\sum_{i\in I_N}Y_i)}$$
$$\times\prod_{n=1}^{N}(Se_n^{\sum_{i\in C_n}Y_i}(1-Se_n)^{\sum_{i\in\bar{C_n}}Y_i}Sp_n^{\sum_{i\in\bar{C_n}}(O_i-Y_i)}(1-Sp_n)^{\sum_{i\in C_n}(O_i-Y_i)}). \tag{6}$$

Hence, our case is a consequence of the more general result that applies to dichotomic diagnostic tests, i.e. the probability model used is the Binomial. In the Bayesian approach, the problem is to know the correct prior distribution. We assume that all test parameters π, Se_n and Sp_n follow Beta(α,β) prior distribution. So by considering conditionally independent tests, the joint multivariate prior distribution is

$$f(\pi,\mathbf{Se},\mathbf{Sp})\propto\text{Beta}(\alpha_\pi,\beta_\pi)\prod_{n=1}^{N}\text{Beta}(\alpha_{Se_n},\beta_{Se_n})\text{Beta}(\alpha_{Sp_n},\beta_{Sp_n}). \tag{7}$$

The posterior distribution f is obtained from Bayes' theorem using the expression (6) of the likelihood function

$$f(\pi,\mathbf{Se},\mathbf{Sp}|\mathbf{O},\mathbf{Y})\propto l(\mathbf{O},\mathbf{Y}|\pi,\mathbf{Se},\mathbf{Sp})f(\pi,\mathbf{Se},\mathbf{Sp}). \tag{8}$$

The joint posterior distribution $f(\pi,\mathbf{Se},\mathbf{Sp}|\mathbf{O},\mathbf{Y})$ is given by the product of $2N+1$ beta distributions, with unknown parameters α and β functions of the observed and the latent data. Indeed, the form of the observational model (6) preserves the form of the posterior with respect to the prior (conjugacy) [8]. From the posterior joint distribution of f and from the distribution (5) of the latent variables, we obtain the following full conditional distributions for all unknown variables

$$Y_i | O_i, \pi, \mathbf{Se}, \mathbf{Sp} \sim \text{Binomial}(O_i, p_i), \tag{9}$$

$$\pi | \mathbf{O}, \mathbf{Y}, \alpha_\pi, \beta_\pi \sim \text{Beta}\left(\sum_{i \in I_N} Y_i + \alpha_\pi, N_\text{s} - \sum_{i \in I_N} Y_i + \beta_\pi\right), \tag{10}$$

$$Se_n | \mathbf{Y}, \alpha_{Se_n}, \beta_{Se_n} \sim \text{Beta}\left(\sum_{i \in C_n} Y_i + \alpha_{Se_n}, \sum_{i \in \bar{C}_n} Y_i + \beta_{Se_n}\right), \tag{11}$$

$$Sp_n | \mathbf{O}, \mathbf{Y}, \alpha_{Sp_n}, \beta_{Sp_n} \sim \text{Beta}\left(\sum_{i \in \bar{C}_n} (O_i - Y_i) + \alpha_{Sp_n}, \sum_{i \in C_n} (O_i - Y_i) + \beta_{Sp_n}\right). \tag{12}$$

With the previous expressions for the full conditional distributions of the $1 + 2N + 2^N$ unknown variables, one for prevalence, $2N$ for sensitivity and specificity and 2^N for the latent variables, it is possible to apply the Gibbs sampler to obtain a sample of each variable drawn from its posterior distribution.

With the previous expressions of the full conditional distributions of the $1 + 2N + 2^N$ unknown variables, one for the prevalence, $2N$ for the sensitivity and the specificity and 2^N for the latent variables, it is possible to apply the Gibbs sampler to obtain a sample of each variable drawn from its posterior distribution.

3. The Gibbs sampler implementation

The Gibbs sampler operates by fixing arbitrary starting values of the unknown quantities $\pi^{(0)}$, $Se_n^{(0)}$ and $Sp_n^{(0)}$, $n = 1, \ldots, N$ and by using the expressions (9)–(12) to draw new values from the full conditional distributions. At the jth iteration the new values are obtained through successive generations by means of the following scheme:

$$Y_i^{(j)} \sim \text{Binomial}(O_i, p_i^{(j-1)}), \tag{13}$$

with $i = 1, \ldots, 2^N$, and

$$\pi^{(j)} \sim \text{Beta}\left(\sum_{i \in I_N} Y_i^{(j-1)} + \alpha_\pi, N_s - \sum_{i \in I_N} Y_i^{(j-1)} + \beta_\pi\right) \tag{14}$$

and

$$Se_n^{(j)} \sim \text{Beta}\left(\sum_{i \in C_n} Y_i^{(j-1)} + \alpha_{Se_n}, \sum_{i \in \bar{C}_n} Y_i^{(j-1)} + \beta_{Se_n}\right) \tag{15}$$

$$Sp_n^{(j)} \sim \text{Beta}\left(\sum_{i \in \bar{C}_n} (O_i - Y_i^{(j-1)}) + \alpha_{Sp_n}, \sum_{i \in C_n} (O_i - Y_i^{(j-1)}) + \beta_{Sp_n}\right), \tag{16}$$

with $n = 1, \ldots, N$.

Thus, at each iteration of the Gibbs sampler, $1 + 2N + 2^N$ random generations are performed from all full conditional distributions. J iterations are performed: the first M iterations ensure the convergence of all posterior full conditional distributions, whereas the remaining $J - M$ are used for inference from posterior distributions. From each value of $(\pi, \mathbf{Se}, \mathbf{Sp})$ of the posterior sample, the positive predictive value (PPV_n) and the negative predictive value (NPV_n) of each test are calculated. Expressed in words, the PPV is defined as the probability that an individual that has tested positive within the diagnostic test is truly diseased, whereas NPV is the probability that an individual that has tested negative is truly free of disease.

Once convergence has been achieved, the information is summarized by calculating the medians and the 0.95 credibility intervals of all variables.

3.1 *Convergence diagnostic*

When implementing the Gibbs sampler, the problem of correctly choosing the values of the number of burn-in iterations M and the number $J - M$ of sub sequential iterations used for inference from posterior distributions arises. The choice of M is related to the convergence of Gibbs sampling. To check the convergence, several methods have been proposed. In [3] a review of the methods used for the assessment of the convergence of MCMC simulations is presented, with an emphasis on the theoretical aspects. Some informal convergence monitoring techniques were proposed in [9] and also presented in [8]. Some of these are based on inspection of the plots of the ergodic average of the chain as a function of the number of iterations of the Gibbs sampler. Asymptotic behavior over many successive iterations indicates convergence. In [18] such a method was used by plotting the ergodic averages of the medians of the prevalence, sensitivity and specificity as a function of the number of iterations.

We implement the Gibbs sampler with a single chain, by assuming the ergodic property of Markov chains [8,11]. This approach is different from those used by the authors of the method, which use several chains processed in parallel [15]. In the software package [14], it is possible to select up to five chains.

Another problem when implementing the Gibbs sampler arises if the autocorrelation of the chain is too high. In this case, the sample obtained is not representative of the entire parameter space, at least if J is not large enough. For this reason, the technique referred to as *thinning* the chain is sometimes used, which involves saving the sample at every kth iteration. The lag k between iterations allows the correlation of the chain to be decreased, although the number of iterations required to obtain the same sample size increases. The software package [14] implements *thinning*, proposing its use when dealing with large sample sizes (e.g. several thousands of subjects).

Here, we use the method proposed by R&L [21,22] to perform convergence diagnostic of the Gibbs sampling presented above. This method is implemented in BOA [1] software, which is an R/S-PLUS program for carrying out convergence diagnostics of Monte-Carlo sampling outputs and also in the Fortran program *gibbsit*, available free of charge at http://lib.stat.cmu.edu. This method provides the criteria for choosing M, k and J_{prec}, the number of iterations successive to the first M of which every k stored is required to obtain the posterior estimates with a known precision. This method works when the quantiles q of interest of the parameters obtained from any MCMC algorithm have to be estimated, as is the case with the Gibbs sampler. In our case, we consider the following qth quantiles as the parameters for the diagnostic tests: $q = 0.50$ for calculating the medians and $q = 0.025$ and 0.975 for calculating the 0.95 confidence intervals.

This method estimates the values of M, k and J_{prec} that are sufficient to obtain q with an error smaller than the required precision $\pm r$ and with a confidence s. The estimated quantile $\hat{q}$ of q therefore satisfies the condition

$$P(|\hat{q} - q| \le r) = s.$$

Table 1. Minimum number of iterations $J_{\min}$ required to estimate the quantile $q = 0.025$ at different values of the precision r, with confidence level $s = 0.95$ [21].

r	0.0025	0.005	0.0075	0.01	0.0125	0.015	0.02
$J_{\min}$	14982	3748	1665	936	600	416	234

For example, if with the Gibbs sampler we want to estimate the quantile $q = 0.025$ with the precision $r = 0.0125$ and with $s = 0.95$, the estimated quantile $\hat{q}$ lies between 0.0125 and 0.0375, with 95% confidence. It should be noted that the precision r refers to the cumulative distribution function at the quantile of interest of the investigated parameters. This precision r should not be confused with the uncertainty of the parameter at the quantile q drawn from the posterior distribution.

We use the *gibbsit* to obtain the appropriate values of M, k and J_{prec}. The method requires the Gibbs sampler to perform a set of initial pilot samples, one for the sensitivity and the specificity of each diagnostic test and one for the prevalence, with q, r and s as input parameters. It returns the values of M, k and J_{prec}. This algorithm also gives the minimum size $J_{\min}$ of the initial pilot sample, i.e. the minimum number of iterations required for the chosen values of q, r and s. The expression of $J_{\min}$ as a function of q, r and s is reported in [21,22]. Table 1 shows the minimum number of iterates $J_{\min}$ required to estimate the quantile $q = 0.025$ with confidence level $s = 0.95$ for some values of r.

In order to check the correlation in the chains, the *gibbsit* outputs two parameters k_{thin} and k_{ind}. The first is the number of values to be skipped in the sample which will be sufficient to produce a first-order Markov chain, i.e. when each value of the chain depends only on the previous ones. The second value is the number of values to be skipped in the sample which will result in independent values in the Markov chain. Therefore one would always expect k_{ind} to be larger than k_{thin}. In any case, when k_{thin} or k_{ind} are greater than one, all J_{prec} iterations can be used for inference, at the expense of reduced computational efficiency.

Besides the above parameters, the *gibbsit* outputs $I = (M + J_{\text{prec}})/J_{\min}$. This statistic measures the increase in the number of iterations due to the dependence in the sequence [21]. Values of I greater than 1 indicate a high level of dependence in the chain, which may be due to bad starting values. In [21] changing implementation in the case of values of I greater than about 5 is suggested.

3.2 *How to choose the precision r?*

When the estimate $\hat{x}_q$ of the posterior quantile x_q of the underlying variable x is calculated with the Gibbs sampler, it is more interesting to evaluate the precision of that estimate, i.e. the difference $\hat{x}_q - x_q$, than the precision of $\hat{q}$. It should again be noted that the precision r refers to the error of the cumulative posterior distribution $F(x)$ of x at the quantile q, i.e. $q = F(x_q)$. Thus, $\hat{q}$ lies in $q \pm r = F(x_{q\pm r})$.

Now, we evaluate how the precision r is related to that of x_q for the case of the beta posterior distributions (8). If r is known, the error in x_q depends on the cumulative distribution of $F(x)$. In [21] this error is evaluated by defining

$$e_{x_q} = \max\left\{\frac{x_q - x_{q-r}}{x_q}, \frac{x_{q+r} - x_q}{x_q}\right\}, \tag{17}$$

and is calculated for the Normal (light-tailed), t_4 (moderate tails) and Cauchy (heavy-tailed) distributions. It is shown that the error increases going from the light to the heavy-tailed distribution.

Moreover, when the posterior distribution is unknown, it is suggested that r be fixed by considering the worst case of heavy-tailed distribution with an error given approximately by that of Cauchy distribution.

Here, we examine in greater depth the dependence of r on the precision of the estimated quantiles in the case of the beta distribution. In our case, the estimate parameter x is the prevalence, the sensitivity and the specificity of each test, which follows the beta distribution with unknown values of α and β. Hence, the relative error e_{x_q}, when q and r are fixed, depends on the values of α and β of the posterior beta distribution of the parameter x. If e_{x_q} is known, we can evaluate with the *gibbsit* the number of iterations of the Gibbs sample which guarantees a relative error in the estimated quantile x_q below e_{x_q}.

In order to calculate the quantity e_{x_q}, we have to deal with the cumulative distribution of the beta probability density, which is the regularized incomplete beta function

$$I_x(\alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1} \frac{\mathrm{d}t}{B(\alpha, \beta)}, \tag{18}$$

where $B(\alpha, \beta)$ is the beta function, and the equation

$$q = I_{x_q}(\alpha, \beta) \tag{19}$$

has to be solved for x_q. This can only be done using numerical methods. An approximate expression for e_{x_q} can be obtained by expanding $I_x(\alpha, \beta)$ at x_q up to the first term, valid when $r \ll q$

$$e_{x_q} \cong \frac{rB(\alpha, \beta)}{x_q^{\alpha}(1-x_q)^{\beta-1}}. \tag{20}$$

It should be noted that for α and β fixed the relative error is higher at the tail quantiles than at the medians. In particular, the minimum occurs when $x_q = \alpha/(\alpha + \beta - 1)$, which is a good approximation of the mean value of the beta distribution when α or β are greater than a few units. Therefore, in the following treatment we will mainly consider the relative error only at the 0.025 and 0.975 quantiles. If we choose q and r and then calculate e_{x_q}, the number of iterations of the Gibbs sampler needed to estimate x_q with an error of less than e_{x_q} can be determined.

Figure 1 shows the values of the relative error e_{x_q}% at the quantile x_q for the beta distribution as a function of α and for $\beta = 1, 4, 20$ and 100, calculated with $r = 0.005$ for the 0.025 (Figure 1(a)) and the 0.975 (Figure 1(b)) quantile. It should be noted that the relative error is greater at the 0.025 than at the 0.975 quantile. In both cases, the error increases as α decreases. For the 0.025 quantile the relative error is less sensitive to β than the relative error at the 0.975 quantile. Figure 1 also shows the values of the relative error calculated with the approximated expression (17). For different values of r the curves of the relative error as a function of α and β show a similar behavior, but with higher values at smaller r.

To evaluate how the values of the precision r affect the relative error, we consider the maximum value of e_{x_q}, calculated over α and β, for different values of r. Table 2 displays these maximum values of e_{x_q}. It should be noted that for $r = 0.02$, which is a value used in the implementations of the R&L method [6], the values of the error at the lower quantile for the investigated model can be up to 81.8%.

The values of α and β of the beta posterior distributions depend on those of the prior distributions, and increase with the sample size N_{S}. Hence, the error in the estimation of the parameters of the diagnostic tests with the Gibbs sampler is higher when dealing with scarcely informative *a priori* distributions and with small sample size.

In order to avoid grossly overestimating the quantity e_{x_q} in the quantile of each parameter, rather than considering the maximum values of Table 2 when dealing with informative priors, it is preferable to calculate e_{x_q} for the values of α and β used in the prior distributions.

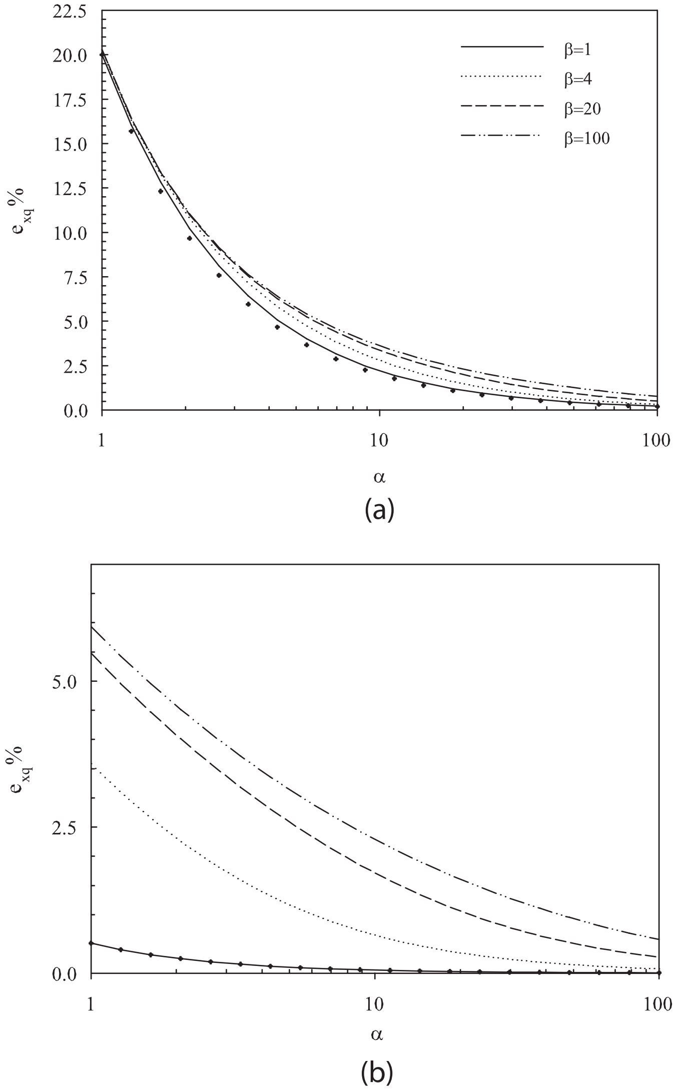


Figure 1. Plots of the relative error e_{x_q}% at the quantile x_q for the beta distribution as a function of α and for $\beta = 1$, 4, 20 and 100, calculated with $r = 0.005$ and (a) with $q = 0.025$ and (b) with $q = 0.975$. The dots (●) represent the values of the relative error calculated with the approximated expression (17).

Table 2. Maximum of the relative percent error of the maximum error e_{x_q} calculated for the quantiles 0.025 and 0.975, for different values of the precision r.

r	$e_{x_{0.025}}$%	$e_{x_{0.975}}$%
0.0025	10.1	2.8
0.0050	20.3	5.9
0.0075	30.5	9.5
0.0100	40.7	13.6
0.0125	50.9	18.4
0.0150	61.2	24.3
0.0200	81.8	42.5

4. Application of the Bayesian analysis with Gibbs sampler to estimate the parameters of the WST in GERD

We consider a data set of 172 GERD patients collected to evaluate the accuracy of the WST associated with a barium study compared with three other frequently used imperfect reference tests, 24 h pH monitoring, endoscopy and histology. Data were collected and provided for the analysis by the Esophageal Surgical Unit of the Department of Surgical Oncology of the University of Palermo.

GERD is one of the most frequent benign diseases of the upper gastrointestinal tract. It develops when the reflux of gastric content causes troublesome symptoms or complications [23]. The atypical symptoms are angina-like pain, chronic cough, asthma, hoarseness, protracted hiccups, globus sensation, dental erosion, ear pain, night sweats and sleep apnea, and water brash. Typical GERD symptoms include regurgitation, heartburn and dysphagia. As GERD is associated with numerous different symptoms, not always specific, all the subjects included in the analysis were symptomatic and the problem was to estimate the diagnostic accuracy of tests used to investigate whether symptoms might be reflux related. It is therefore difficult to detect the disease condition of individuals of one population because clinical and anatomical indicators are not equivalent and must be considered as different tools. For this reason, prevalence is difficult to estimate because only symptomatic patients are investigated.

The four test outcomes for a given subject can be considered conditionally independent because they measure different biological processes. In fact, the 24 h pH monitoring test is generally used for detection of acid gastroesophageal reflux, esophageal endoscopy is usually performed when severe symptoms are present or when a complication is suspected, and histology is performed to confirm esophagitis, or to detect Barrett's esophagus. WST shows reflux as a mechanical event independently of its chemical composition.

We use the following notation to identify one of the four tests used: WST → Test 1, 24 h pH monitoring → Test 2, endoscopy → Test 3, histology → Test 4.

The values of the parameters α and β of the beta prior distributions for the prevalence and for the sensitivity and specificity of each of the four tests are based on expert opinion and a review of the literature [7,17,19,24]. The parameters of the beta prior for each test parameter are computed, using the method of the moments [10,15], by matching the mean of the beta distribution with the mean value $\bar{x}$ of the data obtained from literature and clinical opinion, as well as its standard deviation with one quarter of the range R_x of these data,

$$\alpha = -\frac{\bar{x}(16\bar{x}^2 - 16\bar{x} + R_x^2)}{R_x^2} \tag{21}$$

$$\beta = \frac{(16\bar{x}^2 - 16\bar{x} + R_x^2)(\bar{x} - 1)}{R_x^2}. \tag{22}$$

Table 3 shows the values of the α and β coefficients of the prior distributions. As insufficient information is available in the literature concerning the accuracy of the WST test, we used the non-informative beta prior for it (i.e. $\alpha = 1\beta = 1$).

From the test results, we obtained the vector **O** for the number of subjects which have one of the 16 different test patterns (see Table 4).

We implemented the Gibbs sampler as in Section 2 with $N = 4$ by using the MATLAB$^{6.5}$ software package. We chose the initial values of $\pi^{(0)}$, $Se_n^{(0)}$ and $Sp_n^{(0)}$ along the nine-dimensional unitary cube. In order to choose arbitrary values for these parameters, we used the MATLAB function *rand*, which generates uniformly distributed random numbers in the interval [0,1]. We also performed simulations with initial values of $\pi^{(0)}$, $Se_n^{(0)}$ and $Sp_n^{(0)}$ placed near the corners

Table 3. Prevalence and coefficients of the beta prior distributions for the parameters of GERD diagnostic tests.

Prevalence (GERD)		
Range	0.30–0.50	
Beta coefficients	$\alpha = 38, \beta = 57$	
	Sensitivity	Specificity
Test 1 (WST)		
Range	–	–
Beta coefficients	$\alpha = 1, \beta = 1$	$\alpha = 1, \beta = 1$
Test 2 (24 h pH monitoring)		
Range	0.80–0.90	0.80–0.90
Beta coefficients	$\alpha = 172.6, \beta = 30.45$	$\alpha = 172.6, \beta = 30.45$
Test 3 (endoscopy)		
Range	0.50–0.70	0.70–0.90
Beta coefficients	$\alpha = 57, \beta = 38$	$\alpha = 50.40, \beta = 12.60$
Test 4 (histology)		
Range	0.80–0.99	0.80–0.99
Beta coefficients	$\alpha = 36.38, \beta = 4.268$	$\alpha = 36.38, \beta = 4.268$

Note: A uniform distribution was used for the prior distribution for the WST for GERD.

Table 4. Results of WST (Test 1), 24 h pH monitoring (Test 2), endoscopy (Test 3) and histology (Test 4) tests for GERD on 172 individuals.

Test 1	Test 2	Test 3	Test 4	Number of observations
0	0	0	0	3
0	0	0	1	13
0	0	1	0	2
0	0	1	1	4
0	1	0	0	2
0	1	0	1	12
0	1	1	0	1
0	1	1	1	7
1	0	0	0	4
1	0	0	1	25
1	0	1	0	0
1	0	1	1	24
1	1	0	0	2
1	1	0	1	28
1	1	1	0	11
1	1	1	1	34

of the nine-dimensional unitary cube to test the convergence of the Gibbs sampling more effectively. At each iteration of the Gibbs sampling, we chose the MATLAB function *binornd*, which generates random numbers from the binomial distribution, to implement Equation (13) and the function *betarnd*, which generates random numbers from the beta distribution, to implement Equations (14)–(16). In order to avoid reproducing the same output of the random number functions, we changed the MATLAB variable *state* each time before running simulations.

Firstly, we applied the Gibbs sampler to the data obtained from our four test results and used the prior parameters of Table 3 to generate the pilot samples of the test parameters, one for the prevalence, four for sensitivity and four for specificity. We chose $r = 0.005$ and $s = 0.95$, so the

Table 5. Output parameters of *gibbsit* program for pilot samples for GERD prevalence and the sensitivity and specificity of the four diagnostic tests with the 172 observed data, for the $q = 0.025$ and $q = 0.975$ quantiles and with the precision $r = 0.005$, with informative priors.

	$q = 0.025$					$q = 0.975$				
	k_{thin}	k_{ind}	M	J_{prec}	I	k_{thin}	k_{ind}	M	J_{prec}	I
π	1	2	5	5687	1.52	1	3	4	5105	1.36
Se_1	1	2	4	4778	1.28	1	2	5	5330	1.42
Se_2	1	2	3	4447	1.19	1	2	4	3650	1.24
Se_3	1	2	2	3993	1.07	1	1	2	3555	0.95
Se_4	1	2	2	4025	1.08	1	2	4	4941	1.32
Sp_1	1	2	4	5105	1.36	1	2	4	4882	1.30
Sp_2	1	2	2	3989	1.07	1	2	3	3711	0.99
Sp_3	1	3	4	4826	1.29	1	2	4	4920	1.31
Sp_4	1	2	2	3880	1.04	1	2	4	4667	1.25

relative errors in the estimated test parameters are less than 20% for the 0.025 quantile and less than 6% for the 0.975 quantile (see Table 2).

We generated pilot sequences by running the Gibbs sampler with 4000 iterations, just above the minimum value $J_{\min} = 3746$. By changing the starting values of the Gibbs sampler, new pilot samples were obtained and by running the *gibbsit* again, new values of the parameters k_{thin}, k_{ind}, M, J_{prec} and I were obtained. Table 5 shows the output parameters of *gibbsit* program for the pilot samples of the four diagnostic tests, with $q = 0.025$ and $q = 0.975$, for the case $r = 0.005$, with informative prior distributions. When the starting values are changed, the output parameters do not vary significantly.

It should be noted that the values of M were less than 10 in all cases, so a few burn-in iterations were required to obtain convergence. $k_{\text{thin}} = 1$ in all cases, so the chains for all parameters are first-order Markov chains and k_{ind} is about a few units. The small values of k_{ind} indicate that the correlation in the chains is low. This is also confirmed by the values of the statistic I, which was always close to 1. The number of iterations of the Gibbs sampler required to obtain the required precision J_{prec}, i.e. $e_{x_{0.025}} < 20\%$ and $e_{x_{0.975}} < 6\%$, was never higher than 6000.

We also performed simulations with the same data without informative priors, i.e. by setting $\alpha = \beta = 1$ for π, Se_n and Sp_n, with $n = 1, \ldots, N$. In Table 6 are displayed the values of the output parameters of the *gibbsit*. The results show that, although the number of burn-in iterations

Table 6. Output parameters of *gibbsit* program for pilot samples for GERD prevalence and the sensitivity and specificity of the four diagnostic tests with the 172 observed data, for the $q = 0.025$ and $q = 0.975$ quantiles and with the precision $r = 0.005$, without informative priors.

	$q = 0.025$					$q = 0.975$				
	k_{thin}	k_{ind}	M	J_{prec}	I	k_{thin}	k_{ind}	M	J_{prec}	I
π	2	3	44	46,668	12.47	2	26	8	26,526	7.09
Se_1	7	2	84	100,877	26.95	1	10	4	10,576	2.83
Se_2	2	2	24	28,814	7.70	2	14	4	13,786	3.68
Se_3	5	1	45	37,415	10.00	2	16	8	17,498	4.68
Se_4	2	2	22	26,160	6.99	2	12	5	13,682	3.66
Sp_1	1	2	8	9048	2.42	4	40	13	45,080	12.04
Sp_2	5	30	30	43,365	11.58	1	11	7	11,890	3.18
Sp_3	5	25	25	36,040	9.36	3	21	8	24,699	6.68
Sp_4	2	12	12	14,858	3.97	5	40	12	37,200	9.94

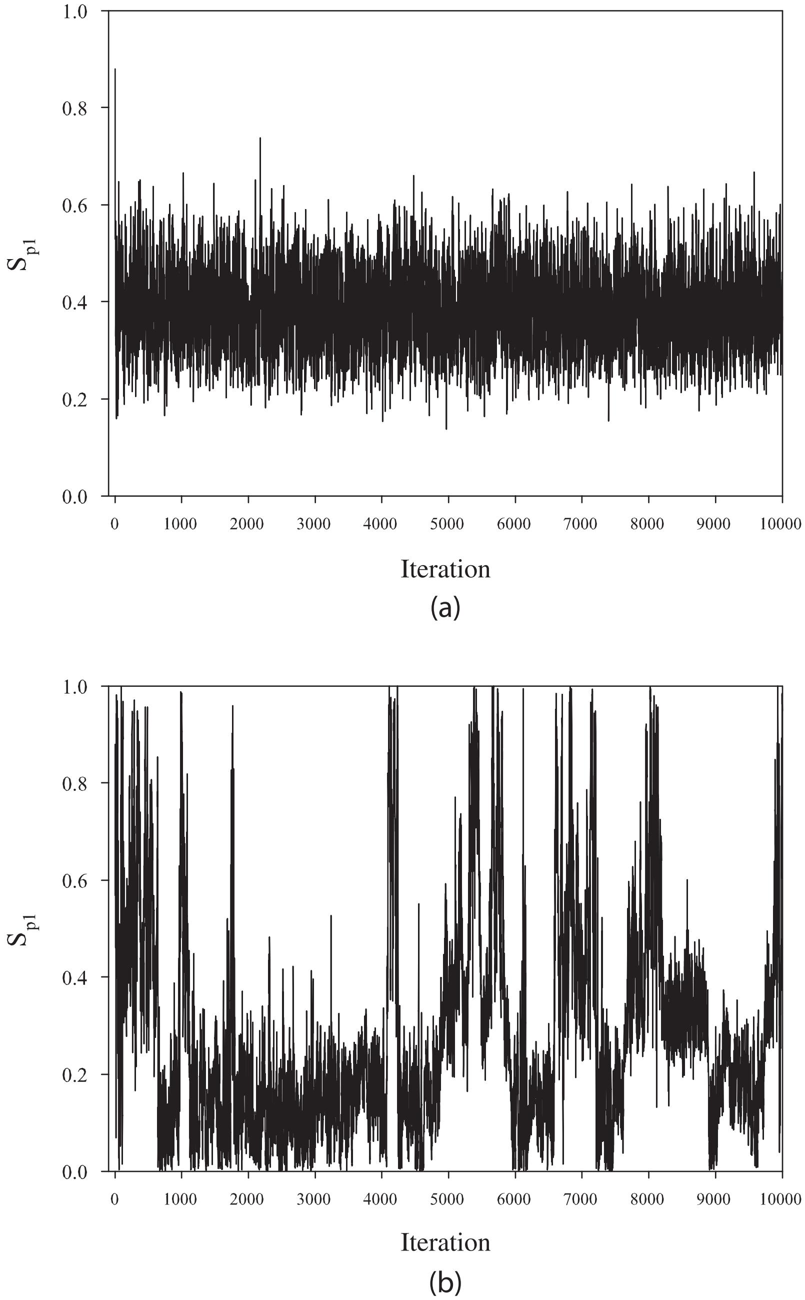
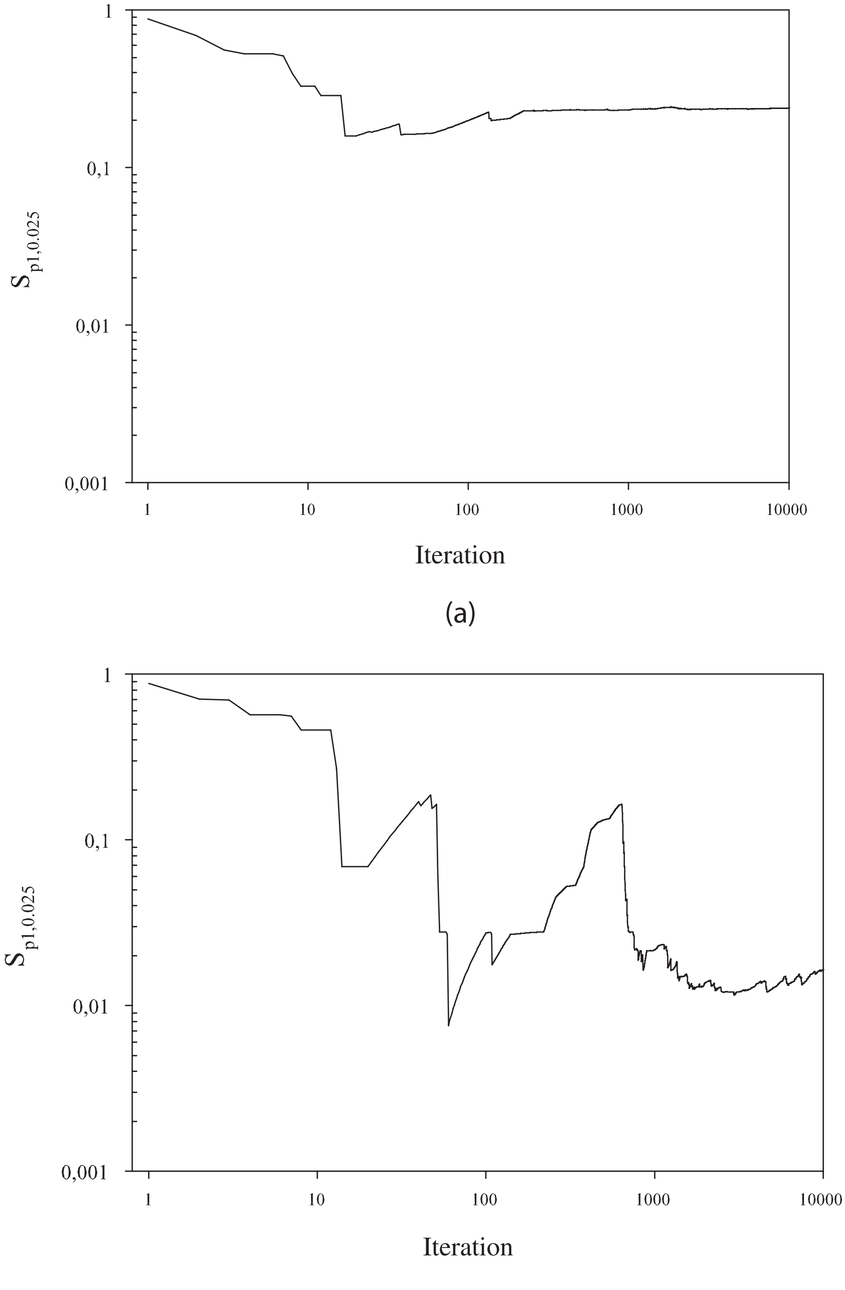


Figure 2. Plots of the WST specificity with number of iterations of the chain of the Gibbs sampler for the 0.025 quantile, (a) with beta and (b) with uniform prior distribution.

M increases by roughly only an order of magnitude, a large number of iterations, up to 100,000, is required to obtain the same precision $r = 0.005$ for both considered quantiles. Moreover, the values k_{thin} and k_{ind} in all cases are greater than 1, so the correlation in the chains of the parameters is high. Finally, it should be noted that the statistic I mostly assumes values greater than 5. This indicates poor implementation of the Gibbs sampler [22]. By varying the starting values to generate pilot samples, the *gibbsit* generates similar large values.

The difference in convergence between the case of informative and non-informative prior distributions can be also be illustrated graphically. For example, the values of specificity of the



(a)

(b)

Figure 3. Log–log plots of ergodic averages of the WST specificity with number of iterations of the chain of the Gibbs sampler for the 0.025 quantile, (a) with beta and (b) with uniform prior distribution.

Table 7. Medians and 0.95 credible intervals of the output parameters of the accuracy of the water-siphon (Test 1), 24 h pH monitoring (Test 2), endoscopy (Test 3) and histology (Test 4) tests for GERD for the 172 investigated subjects.

	Prevalence 0.53 [0.45–0.62]			
	Test 1	Test 2	Test 3	Test 4
Sensitivity	0.82 [0.71–0.90]	0.83 [0.78–0.88]	0.60 [0.53–0.68]	0.90 [0.83–0.95]
Specificity	0.38 [0.23–0.55]	0.84 [0.78–0.88]	0.76 [0.66–0.84]	0.46 [0.36–0.58]
NPV[a]	0.65 [0.45–0.81]	0.81 [0.72–0.88]	0.62 [0.53–0.72]	0.80 [0.68–0.89]
PPV[b]	0.60 [0.49–0.72]	0.85 [0.78–0.91]	0.74 [0.62–0.84]	0.66 [0.55–0.76]

Note: The results are obtained with 10 burn-in and 6000 sub-sequential iterations of the Gibbs sample.
[a] Negative predictive value.
[b] Positive predictive value.

WST generated by the Gibbs sampler are shown in Figure 2 as a function of the iterations in a run of the chain. Figure 2(a) shows those with informative prior distributions, whereas Figure 2(b) refers to a run with $\alpha = \beta = 1$ in all prior distributions (i.e. uniform distributions). The starting value is 0.88 for both chains. It should be noted that the first chain shows good convergence. Indeed, the white noise behavior of this series indicates low correlation in the chain and the stationary in the variance ensures convergence of the quantile within a few Gibbs iterates. Conversely, in the case of uniform prior distributions for all diagnostic test parameters, the same series is more highly correlated and its variance is non-stationary.

Figure 3 shows the ergodic averages of the runs of Figure 2 in a log–log plot. It should be noted that in the first case (Figure 3(a)), after a short initial transient period the 0.025 quantile exhibits asymptotic behavior toward 0.23 (see Table 7). In the case of non-informative prior distributions (Figure 3(b)), note that the initial burn-in period of $M = 8$ iterates given by *gibbsit* (see Table 6)) is followed by almost periodic oscillations which relax after one thousand iterates, over which the average does not exhibit asymptotic behavior. These oscillations do not alter the average final value (≈0.016). This explains the low value of M obtained but indicates convergence problems.

Moreover, the low value of 0.016 for the 0.25 quantile of the WST specificity in the implementation without informative priors demonstrates that the 0.95 confidence interval limits of the specificity of the WST are too broad. The same wide confidence intervals are obtained for the other parameters of the other diagnostic tests in the case of non-informative priors.

So in order to obtain an estimate of the prevalence of GERD and that of the parameters of the four tests used, we ran the Gibbs sampling with $M = 10$ and $J_{\text{prec}} = 6000$. We used all these iterates for inference because $k_{\text{thin}} = 1$ for all parameters. Table 7 shows the values of the medians and the 95% credibility intervals, obtained from posterior distributions, for the prevalence of GERD, the sensitivity, specificity, PPV and the NPV of the four diagnostic tests from the data set of the 172 investigated subjects.

5. Discussion

In this study, we used the Bayesian analysis for multiple binary diagnostic test evaluation in the absence of a gold standard. The latent data and observations from the joint posterior were simulated in the Bayesian approach by an iterative MCMC technique using the Gibbs sampler. The simulated samples were then used to approximate the actual posterior distributions.

In the attempt to use this Gibbs sampler algorithm to evaluate the performance of the WST, and of 24 h pH monitoring, endoscopy and histology, none of which can be considered a gold standard, we investigated how its implementation affects the accuracy of the parameters of these tests. Indeed, incorrect implementation of the Gibbs sampler can affect the accuracy of the parameters of the diagnostic tests obtained from posterior distributions. For this purpose, we applied the method of R&L [21,22], valid for investigating the convergence of MCMC algorithms, to determine the error in the estimated quantiles of the posterior distributions with a known accuracy.

One of the main results of our work is that we obtained a method for checking the accuracy of the estimated quantiles of the posterior distributions obtained with the Gibbs sampler in the case of N diagnostic independent tests and one population. Indeed, in [22] the error in the 0.025 quantile was evaluated only in the case of Normal, t_4 and Cauchy distributions. When the behavior of the tails of the posterior distribution is unknown in advance, the authors of the method suggest fixing r and then $J_{\min}$ by considering the worst case of the Cauchy distribution. Here we show how this error can be calculated and derive an approximate expression for it, which gives this error when the parameters α and β are known. The relation of the quantile error with the parameters of the posterior beta distribution can be used to select an appropriate number of iterations of the Gibbs sampler.

As described in Section 3.2, we applied the *gibbsit* only to the tail quantiles 0.025 and 0.975. This is suggested in [22] for a general case but without any proof of this assertion being provided. We demonstrate this by considering the expression for the error (17).

In [15] the method was applied to two diagnostic tests by performing 20,000 iterations for inference after 500 burn-in iterations, but the authors do not specify how the number of these iterations relates to the accuracy of the estimated test parameters. It should also be noted that in our case the number of burn-in iterations M differs from the number used in [15]. In the software package [14] it is possible to select the values of M and J_{prec}, but the criterion for choosing these values is not given.

We have demonstrated that to obtain the same precision in the estimated parameters for the investigated case, the number of iterations of the Gibbs sampler varies from the case of non-informative to that of informative *a priori*. In the case of non-informative priors, the implementation exhibits several convergence problems which discourage implementation. So when adopting this Gibbs sampler scheme to obtain the parameters of N-diagnostic tests, it is not possible to provide a general rule for determining the appropriate number of iterations in advance.

Evaluation of the error (17) using the method presented above may be useful when implementing the Gibbs sampler because it allows the appropriate number of Gibbs iterates to be selected. We demonstrated that the relative error may be high for small values of α. This generally occurs when the number of observed subjects is small and when non-informative prior distributions are used. To demonstrate this, Table 8 shows the estimated values of α and β of the posterior beta distributions for each parameter of the tests investigated, with the relative errors of each diagnostic test parameter. It should be noted that the relative error is less than 1% in the majority of cases, thus confirming correct implementation of the method for the GERD diagnostic tests with the 172 observed data and with informative priors.

From the above considerations, we can conclude that correct implementation of the Gibbs sampler to N diagnostic tests and one population is critical, and depends strongly on the variance of the prior distributions. It is therefore not possible to give a general rule for determining in advance the number of burn-in iterations and of sub-sequential iterations needed to summarize the information by estimating the quantiles of prevalence, sensitivity and specificity with known precision.

Knowledge of the diagnostic accuracy of a test is important for assessing the efficacy of diagnostic tests in medicine, especially when the test is invasive and other non-invasive tests offering a good level of accuracy are available, at least in the first stages of screening procedures.

Various problems may occur in the context of this complex study given the absence of a gold standard diagnostic test for GERD. One problem concerns the influence of spectrum and selection bias. For example, the judgement of the clinician who has to decide which patients are to undergo diagnostic tests will be crucial. Patients who are referred to hospital tend to have more severe

Table 8. Estimated values of α and β for the posterior beta distributions of the parameters of the four investigated diagnostic tests with informative priors.

	α	β	$e_{x_{0.025}}\%$	$e_{x_{0.5}}\%$	$e_{x_{0.975}}\%$
π	143	124	0.60	0.07	0.47
Se_1	87	20	0.57	0.06	0.31
Se_2	256	52	0.29	0.03	0.20
Se_3	121	79	0.62	0.07	0.46
Se_4	130	15	0.36	0.03	0.18
Sp_1	26	43	1.80	0.19	1.15
Sp_2	226	44	0.30	0.03	0.20
Sp_3	98	32	0.58	0.06	0.36
Sp_4	50	57	1.15	0.13	0.79

Note: The relative errors due to the limited number of Gibbs sampler iterations are for the 0.025, 0.5 and 0.975 quantiles.

symptoms. It is usually easier to discriminate between ‘disease’ and ‘no disease’ if patients have more extreme manifestations of the disorder. Patients with more severe symptoms are referred to secondary or tertiary care and evaluation in this setting will increase the sensitivity and will reduce the specificity of tests in the diagnosis of GERD. It is important to be aware of the influence of spectrum and selection bias on the accuracy of diagnostic tests in GERD, but this must be kept in perspective. It is better to carefully assess the sensitivity and specificity of tests in a selected group than not to do this [19].

Another problem concerns the assumption of conditional independence. In this study, conditional independence has been assumed, in other words the tests must be conditionally independent given the disease status, i.e. the probability of any test outcome given that the disease status is constant across all outcome categories of the other test. Because it simplifies the statistical problem, numerous methods have been developed on the basis of the assumption of conditional independence [12] in spite of the fact that in some practical situations it is not realistic, e.g. when there is a spectrum of disease severity that may induce correlation between the tests. All tests may be positive in subjects with severe disease, whereas false negatives may be more likely in subjects with mild disease [20]. A different situation occurs when two or more diagnostic tests may be conditionally dependent due to a factor other than the disease status, for example arising from a common biological phenomenon on which the tests are based. If the test is used in combination with other tests whose performance is affected by the severity of the disease, then dependence would be induced between the tests [5]. There is no agreement in the literature about the way to assess the validity of the conditional independence assumption, as it is difficult to verify in practice [16]. The conditional independence assumption can be directly tested only when data on disease are available and, if it is in doubt, a dependence structure must be specified [20]. An alternative approach is to assume conditional independence (i.e. the conditional correlations are zero), providing that a biological argument can be made to support this assertion [2]. In our study of the GERD disease, the assumption of conditional independence between the four test outcomes for a given subject is reasonable because the four tests measure different biological processes.

As regards the results of simulations, Table 7 gives the medians and of the 0.95 credibility intervals of the prevalence of GERD, the sensitivity and the specificity of the 172 investigated subjects. The prevalence of GERD is estimated to be 0.53 with [0.45–0.62]. In the application of the Bayesian analysis with Gibbs sampler, we focus mainly on the performance of the WST because the sensitivity and the specificity of this test is scarcely documented with respect to the other three tests, which are also used in the screening of different diseases.

We obtain for the sensitivity and the specificity of the WST 0.82 [0.71–0.90] and 0.38 [0.23–0.55], respectively.

Simulation results (Table 7) show that the WST has a high sensitivity the histology. One possible reason is that reflux may cause histologically detectable microscopic damage that is not yet detectable endoscopically. WST evidences reflux as a mechanical event independently of its chemical composition and the occurrence of complications. The low sensitivity of endoscopy can be explained because only GERD complications can be detected. This is the reason for high specificity of endoscopy and the low specificity of both WST and histology. As regards the low specificity of histology, the microscopic tissue damage can be caused by minimal reflux episodes that are not yet detectable by other tests.

For the PPV and NPV of the WST we have 0.60 [0.49–0.72] and 0.65 [0.45–0.81], respectively (see Table 7). In our case, the NPV is the more relevant parameter. In practical applications of tests, PPV and NPV are important performance measurements. Indeed, both parameters depend critically on the prevalence of the diagnostic trait, in our case the GERD. A comprehensive discussion of the role of the prevalence is beyond the scope of this work. Briefly, NPV increases if prevalence declines and vice versa. It can be seen that NPV is 100% if the prevalence is zero. A 100% NPV value would be desirable as it indicates that all individuals tested negative are actually non-diseased. The lower the NPV, the higher the risk that truly diseased individuals will test negative.

From the results obtained in our analysis, we found that S_e and NPV of the WST are higher than those of endoscopy, which is an invasive test. We therefore believe that the WST can be considered a valid test for GERD. However, it is important to bear in mind that both sensitivity and specificity parameters should be analyzed to illustrate the disadvantages of accepting tests without considering specificity.

Separate analysis of all parameters in the Table 7 will provide a decision basis for the clinical risk based on the criterion of choice and at the same time will allow specificity to be considered as a criterion on its own, particularly in the first stage of the procedure of screening for GERD. Overall, these results are difficult to analyze from an epidemiology standpoint. Generally speaking, the specificity estimates obtained from the analysis are very low compared with what is expected in routine diagnosis and what has been regularly observed for years by the experts. There may therefore be some bias in our study that could lead to underestimation of the specificity of the tests evaluated. However, the uncertainty regarding the specificity of the WST can be resolved simply by collecting and analyzing more data, although increase in sample size does not guarantee an improvement in the inference of the accuracy of the test parameters [13]. As far as WST is concerned, nothing in the results of the study provides any argument for rejecting this test as suitable for GERD control, as the levels of sensitivity and specificity are no different from those of 24 h pH monitoring, endoscopy and histology. The level of sensitivity obtained with these tests demonstrates that the WST has the capacity to detect diseased individuals with at least the same level of confidence as the other tests that are normally used.

6. Conclusion

In this work we presented a study of the implementation of the Gibbs sampler used to estimate the parameters of N binary diagnostic tests in one population without a gold standard. After the extension of the method proposed by Joseph *et al.* [15] to the case of N diagnostic tests, we investigated the computational aspects of implementation of the Gibbs algorithm of this method.

We used the R&L diagnostic convergence method [21,22] to evaluate the number of burn-in and sub-sequential iterations needed to estimate the quantiles of interest from the posterior beta distributions with a known precision, by calculating the relation for the error in the estimate quantile for the case of the beta distribution as a function of the number of Gibbs iterates. By applying this procedure to the case of four diagnostic tests for GERD, we find that implementation of this method is highly sensitive to the parameters of both the prior distributions and the data. Convergence diagnostic is therefore generally advised for this method.

In the proposed procedure, we evaluated the accuracy of the WST compared to three other imperfect reference tests that are normally used in the GERD diagnostic, i.e. 24 h pH monitoring, endoscopy and histology. For this case we obtained good convergence of the Gibbs sampling and very small errors in the estimated quantiles.

The values for the WST obtained with our analysis compared with those of the other three tests demonstrate that the WST has the ability to detect diseased individuals with at least the same degree of confidence as the other tests. Further investigations should be performed in order to consider in greater detail some critical aspects that affect the accuracy of the tests, such as the planning sampling of the data and the estimate of the effective prevalence of GERD.

Acknowledgements

The authors wish to thank Professor Eugenio Fiorentino, Head of Oesophageal Surgical Unit of the Department of Surgical Oncology of the University of Palermo, for supplying the data for this study. This work was carried out with the partial support of the University of Palermo through the ex-60% funds.

References

[1] N. Best, K. Cowles, and K. Vines, *Bayesian Output Analysis Program BOA*, Department of Biostatistics, The University of Lowa College of the Public Health, 2005. Available at http://www.publichealth. uiowa.edu/boa.

[2] A.J. Branscum, I.A. Gardner, and W.O. Johnson, *Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling*, Prev. Vet. Med. 68 (2005), pp. 145–163.

[3] S. Brooks and G.O. Roberts, *Convergence assessment techniques for Markov chain Monte Carlo*, Stat. Comput. 8 (1982), pp. 319–335.

[4] A. David and A. Skene, *Maximum likelihood estimation of observer error rates using the EM algorithm*, Appl. Stat. 41(28) (1979), pp. 20–28.

[5] N. Dendukuri and L. Joseph, *Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests*, Biometrics 57(1) (2001), pp. 158–167.

[6] M.G. Dodds and P. Vicini, *Assessing convergence of Markov chain Monte Carlo simulations in hierarchical Bayesian models for population pharmacokinetics*, Ann. Biomed. Eng. 32(9) (2004), pp. 1300–1313.

[7] T. Eubanks, P. Omelanczuk, A. Hillel, N. Maronian, C. Popo, and C. Pellegrini, *Pharyngeal pH measurements in patients with respiratory symptoms before and during proton pump inhibitor therapy*, Am. J. Surg. 181(4) (2001), pp. 466–470.

[8] D. Gamerman, *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*, Chapman & Hall/CRC, USA, 1997.

[9] A. Gelfand, S. Hills, A. Racine-Poon, and A. Smith, *Illustration of Bayesian inference in normal data models using Gibbs sampling*, J. Am. Stat. Assoc. 85 (1990), pp. 972–985.

[10] A. Gelman, S. Lewis, H. Stern, and D. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, FL, 2004.

[11] C. Geyer, *Practical Markov chain Monte Carlo with applications to ancestral inference*, J. Am. Stat. Assoc. 90 (1992), pp. 909–920.

[12] S.L. Hui and X.H. Zhou, *Evaluation of diagnostic tests without gold standards*, Stat. Methods Med. Res. 7(4) (1998), pp. 354–370.

[13] W.O. Johnson, J.L. Gastwirth, and L.M. Pearson, *Screening without a gold standard: the Hui–Walter paradigm revisited*, Am. J. Epidemiol. 153(9) (2001), pp. 921–924.

[14] L. Joseph, (2007). Available at http://www.medicine.mcgill.ca/epidemiology/Joseph/Bayesian-Software-Diagnostic-Testing.html.

[15] L. Joseph, T. Gyorkos, and L. Coupal, *Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard*, Am. J. Epidemiol. 141 (1995), pp. 263–272.

[16] M. Ladouceur, E. Rahme, C. Pineau, and L. Joseph, *Robustness of prevalence estimates derived from misclassified data from administrative databases*, Biometrics 63(1) (2006), pp. 272–279.

[17] G. Locke, N. Talley, S. Fett, A. Zinmeister, and A. Melton, *Prevalence and clinical spectrum of gastroesophageal reflux: a population-based study in Olmsted county, Minnesota*, Gastroenterology 112 (1997), pp. 1448–1456.

[18] D. Matranga, A. Vullo, and F. Principato, *Bayesian analysis of diagnostic accuracy for reflux disease in the absence of gold standard*, Proceedings of the Sismec. Cleup (Padova), Monreale (Palermo), 2007, pp. 549–554.

[19] P. Moayyedi, J. Duffy, and B. Delaney, *New approaches to enhance the accuracy of the diagnosis of reflux disease*, Gut 53(4) (2004), pp. 55–57.

[20] M. Pepe and H. Janes, *Insights into latent class analysis of diagnostic test performance*, Biostatistics 8(2) (2007), pp. 474–484.

[21] A. Raftery and S. Lewis, *How many iterations in the Gibbs sampler?* in *Bayesian Statistics*, Oxford University Press, London, 1992, pp. 763–773.

[22] A. Raftery and S. Lewis, *Implementing MCMC*, in *Markov Chain and Monte Carlo in Practice*, Oxford University Press, London, 1992, pp. 115–130.

[23] J. Richter, *Gastroesophageal reflux disease*, Best Pract. Res. Clin. Gastroenterol. 21 (2007), pp. 609–631.

[24] J. Ronkainen, P. Aro, T. Storskrubb, S.-E. Johansson, T. Lind, E. Bolling-Sternevald, H. Graffner, M. Vieth, M. Stolte, L. Engstrand, N.J. Talley, and L. Agréus, *High prevalence of gastroesophageal reflux symptoms and esophagitis with or without symptoms in the general adult Swedish population: a kalixanda study report*, Scan. J. Gastroenterol. 40 (2005), pp. 275–285.

[25] P. Valenstein, *Evaluating diagnostic tests with imperfect standards*, Am. J. Clin. Pathol. 93 (1990), pp. 252–258.

[26] S. Walter and L. Irwig, *Estimation of test error rates, disease prevalences, and relative risk from misclassified data: a review*, J. Clin. Epidemiol. 41 (1988), pp. 923–937.